# Supplementary Material: Anatomy-Constrained Duplication Loss for Tooth Detection

Noam Delbari[1,2][0009−0007−5523−7362], Ariel Hirschhorn[3][0009−0005−2757−6397], and Bella Specktor-Fadida[2][0000−0002−5668−0303]

[1] School of Computer Science, Reichman University, Herzliya, Israel
[2] Department of Medical Imaging Sciences, Faculty of Social Welfare and Health Sciences, University of Haifa, Haifa, Israel
[3] Department of Cranio-Maxillofacial Surgery, Sheba Medical Center, Tel Hashomer, Ramat Gan, Israel

## 1 Gradient Behaviour and Derivation

The detector head and its BCE classification loss are unchanged. The row-wise softmax is used only inside $\mathcal{L}_{\text{dup}}$; inference continues to use the detector's sigmoid confidences. For representative-anchor logits $s_{ij}$, let $p_{ij}$ be the row-wise softmax probability and let $q_j = \sum_i p_{ij}$. The duplication loss is

$$\mathcal{L}_{\text{dup}} = \sum_{j=1}^{C} \max(0,\ q_j - 1)\,. \tag{1}$$

*Derivation.* The loss depends on the logits only through the soft counts $q_j$. The outer hinge contributes

$$\frac{\partial \mathcal{L}_{\text{dup}}}{\partial q_j} = \mathbf{1}[q_j > 1], \tag{2}$$

and the softmax Jacobian is $\partial p_{ij}/\partial s_{ik} = p_{ij}(\delta_{jk} - p_{ik})$. With $\mathcal{A} = \{j : q_j > 1\}$ and $m_i = \sum_{j\in\mathcal{A}} p_{ij}$, the chain rule gives

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{dup}}}{\partial s_{ik}} &= \sum_{j=1}^{C} \mathbf{1}[q_j > 1]\frac{\partial p_{ij}}{\partial s_{ik}} = \sum_{j\in\mathcal{A}} p_{ij}\big(\delta_{jk} - p_{ik}\big) \\ &= \mathbf{1}[k \in \mathcal{A}]p_{ik} - p_{ik}\sum_{j\in\mathcal{A}} p_{ij} = p_{ik}\big(\mathbf{1}[k \in \mathcal{A}] - m_i\big).\end{aligned} \tag{3}$$

Summing Eq. 3 over classes gives

$$\sum_k \frac{\partial \mathcal{L}_{\text{dup}}}{\partial s_{ik}} = m_i - m_i \sum_k p_{ik} = 0. \tag{4}$$

Thus each anchor's gradient is a mass-preserving redistribution: it moves class probability away from over-budget classes and toward the remaining classes without directly suppressing objectness. If no class is over budget, then $\mathcal{A} = \varnothing$ and the gradient vanishes. This redistribution can change a predicted FDI class, but the count constraint alone does not identify the correct replacement label.

**Table 1.** Verification of the closed-form duplication-loss gradient against automatic differentiation.

| Dataset | Seeds | $\max \lvert\text{autograd} - \text{closed}\rvert$ | $\max \lvert\text{row sum}\rvert$ | $\overline{\lVert g \rVert}$ dup. | $\overline{\lVert g \rVert}$ clean |
|---|---|---|---|---|---|
| DENTEX | 5 | $1.79 \times 10^{-7}$ | $2.93 \times 10^{-7}$ | 0.117 | 0.000 |
| HITL | 5 | $1.79 \times 10^{-7}$ | $2.98 \times 10^{-7}$ | 0.227 | 0.000 |

*Empirical verification.* For each validation radiograph, we selected one representative anchor per ground-truth tooth. We use the highest-alignment TAL anchor as a ground-truth-conditioned training proxy for each object; it is not guaranteed to be the box retained at inference. We formed $\mathcal{L}_{\text{dup}}$ and compared Eq. 3 with `torch.autograd` on five converged checkpoints per dataset (image size 1024, rectangular batching disabled). The maximum absolute per-anchor row sum verifies mass preservation, while the duplicate and clean image norms show that the term is active only when a class is over budget.

## 2 Complete Significance Testing

All tests are two-sided paired $t$-tests over the same five seeds (42, 123, 456, 789, 2024); $n = 5$ throughout, so the analysis is underpowered and we report exact $p$-values rather than significance labels alone. Table 2 reproduces every $p$-value behind the resolution sweep of the main paper, including the resolutions and metrics that are not printed there. Table 3 reports the post-processing comparison. The two tables use different evaluators — the per-resolution evaluator and the common evaluator, respectively — so their 1024 px duplicate tests differ in the third decimal (0.028 against 0.026) while agreeing on direction and significance.

Two categories of comparison admit no $p$-value and are omitted rather than reported as non-significant: the reference detectors (DETR, Mask R-CNN, Faster R-CNN) are separate architectures rather than an arm of the baseline/$+\mathcal{L}_{\text{dup}}$ ablation, so although they were also run over the same five seeds there is no paired counterpart to test against; and any duplicate comparison in which both arms are exactly 0.00 has zero variance in the paired difference.

The final block is the honest answer to the question of what the loss adds on top of deterministic deduplication: the mean F1 advantage of 0.2 (HITL) and 0.6 (DENTEX) points is directional only and reaches significance on neither dataset at $n = 5$, as is the recall the loss returns under the same deduplication ($p$=0.090 and $p$=0.071).

## 3 Metric and Evaluator Details

*F1 and duplicate count.* The two headline metrics measure different failure modes. F1 is determined by matched true positives, false positives, and false negatives, whereas the duplicate count records repeated FDI numbers among

**Table 2.** Paired $p$-values for $+\mathcal{L}_{\text{dup}}$ against baseline at every evaluated resolution (per-resolution evaluator, the source of main Table 1). Bold marks $p < 0.05$; ↓ marks a metric on which lower is better. The 640 px and 1024 px rows are those printed in the main paper.

| Dataset | Res | mAP50 | F1 | P | R | Dups↓ |
|---|---|---|---|---|---|---|
| HITL | 512 | 0.069 | 0.259 | 0.067 | 0.717 | 0.118 |
| | 640 | **0.006** | **0.008** | **0.014** | 0.235 | **0.009** |
| | 768 | **0.046** | 0.054 | 0.088 | 0.084 | 0.069 |
| | 1024 | 0.521 | 0.211 | 0.438 | 0.284 | 0.301 |
| DENTEX | 512 | 0.416 | 0.809 | 0.378 | 0.711 | 0.504 |
| | 640 | **0.013** | 0.145 | 0.098 | 0.707 | 0.089 |
| | 768 | 0.143 | 0.073 | 0.158 | 0.157 | 0.125 |
| | 1024 | 0.688 | **0.003** | **0.004** | 0.419 | **0.028** |

predictions surviving NMS. A repeated number is often a false positive, but the relation is not one-to-one: removing a redundant prediction can lower duplicates without changing matched detections, while a better-localized or better-classified box can raise F1 without changing the number of repeated classes. We therefore report both endpoints and do not infer individual relabelling from either aggregate.

*Reference-model input sizing.* The Detectron2 references in Table 1 of the main paper resize by short side to 800 px, with the long side capped at 1600, rather than to a square input as the YOLOv11 arms do. Their numbers therefore provide context and are not an arm of the baseline / $+\mathcal{L}_{\text{dup}}$ comparison.

*Evaluator reconciliation.* Every row of Table 2 of the main paper is scored by one common evaluator, whereas Table 1 uses the per-resolution evaluator. The two agree except in the last printed digit of the duplicate counts (1.32 against 1.31 on HITL, 1.93 against 1.92 on DENTEX), and no comparison changes.

## 4 Matched-Recall Sensitivity Analysis

As a sensitivity analysis separate from the primary fixed-confidence evaluation, we interpolate raw duplicate counts against post-deduplication recall at 1024 px. Each seed is first averaged within a recall band, and the paired test is then performed on those five seed means. The recall bands define only this sensitivity analysis; neither a grid value nor a band is the primary evaluation setting, and confidence 0.25 remains the primary operating point.

**Table 3.** Paired $p$-values for the post-processing comparison of main Table 2 (common evaluator, 1024 px, confidence 0.25). Each row is tested against the untreated Baseline row; bold marks $p < 0.05$. The final block instead tests the loss against deduplication alone, which is the comparison that isolates what the loss adds once a hard one-per-class rule is already applied.

| Dataset | Method (vs. Baseline) | P | R | F1 | Dups↓ | Wrong #↓ |
|---|---|---|---|---|---|---|
| HITL | Soft-NMS | **0.001** | **0.039** | **0.002** | **0.000** | 0.178 |
| | tuned class-aware NMS | **0.002** | **0.004** | **0.002** | **0.001** | 0.178 |
| | Dedup / Hungarian | **0.000** | **0.001** | **0.001** | **0.000** | **0.007** |
| | $+\mathcal{L}_{\text{dup}}$ | 0.454 | 0.307 | 0.241 | 0.308 | 0.095 |
| | $+\mathcal{L}_{\text{dup}}$ + Dedup | **0.000** | 0.250 | **0.001** | **0.000** | **0.013** |
| DENTEX | Soft-NMS | **0.001** | **0.002** | **0.001** | **0.000** | 0.066 |
| | tuned class-aware NMS | **0.001** | **0.027** | **0.001** | **0.000** | 0.540 |
| | Dedup / Hungarian | **0.001** | **0.000** | **0.003** | **0.001** | **0.003** |
| | $+\mathcal{L}_{\text{dup}}$ | **0.005** | 0.389 | **0.002** | **0.026** | **0.024** |
| | $+\mathcal{L}_{\text{dup}}$ + Dedup | **0.000** | 0.071 | **0.000** | **0.001** | **0.004** |
| $+\mathcal{L}_{\text{dup}}$ + *Dedup vs. Dedup alone (duplicates are* 0.00 *in both arms)* | | | | | | |
| | HITL | 0.576 | 0.090 | 0.371 | — | 0.859 |
| | DENTEX | 0.445 | 0.071 | 0.170 | — | 0.118 |

**Table 4.** Matched-recall sensitivity analysis; each seed contributes one band-averaged value. Lower-direction counts and paired $p$-values compare $+\mathcal{L}_{\text{dup}}$ with baseline over the five seed means; bold marks $p < 0.05$.

| Recall band | Dataset | Baseline | $+\mathcal{L}_{\text{dup}}$ | $\Delta$ | Direction; paired $p$ |
|---|---|---|---|---|---|
| 0.85–0.95 | HITL | 1.075 | 0.675 | −0.400 | 5/5 lower; 0.100 |
| | DENTEX | 0.621 | 0.306 | −0.315 | 5/5 lower; **0.005** |
| 0.30–0.95 | HITL | 0.183 | 0.116 | −0.067 | 5/5 lower; 0.075 |
| | DENTEX | 0.108 | 0.055 | −0.053 | 5/5 lower; **0.005** |

## 5 Post-Processing Geometry

We examine the baseline duplicate-pair geometry at the same fixed operating point used for the post-processing comparison. Same-number pairs with sufficient overlap are removable by tuned class-aware NMS; the remainder are NMS-blind.

On these two primary datasets, this geometry explains why tuned NMS removes only about one fifth of baseline duplicate pairs: most same-number boxes do not overlap enough to be suppressed. This is a dataset-specific explanation at the reported operating point, not a claim that NMS is universally ineffective.

## 6 Missing Dentition in Evaluation

This audit demonstrates that missing dentition is present in evaluation; it does not show that the loss solves missing-tooth errors.

**Table 5.** Baseline duplicate-pair geometry at confidence 0.25 and 1024 px, mean over 5 seeds. NMS-removable pairs satisfy the tuned class-aware NMS overlap criterion.

| Dataset | Duplicate pairs | NMS-removable | NMS-blind | Median IoU |
|---|---|---|---|---|
| HITL | 124.6 | 21.6 (18.2%) | 103.0 | 0.133 |
| DENTEX | 218.2 | 41.6 (19.2%) | 176.6 | 0.227 |

**Table 6.** Prevalence of missing dentition in the validation splits.

| Dataset | Validation images | Images with $< 32$ FDI positions | Fraction |
|---|---|---|---|
| HITL | 89 | 73 | 82.0% |
| DENTEX | 96 | 81 | 84.4% |

## 7 Supernumerary Teeth and the Uniqueness Assumption

Neither schema has a dedicated supernumerary class. The DENTEX flags mix potential extra teeth with annotation duplication, so these labels cannot supervise or fairly evaluate the proposed $u_j = 2$ extension.

## 8 Execution-Backed Reproducibility

We audited the fixed 20-run matrix (two datasets, baseline and treatment, five seeds) from checkpoint `train_args` and `results.csv`. The implementation logs the duplication loss under the configuration key `anatomy`, which is therefore the name that appears in the recorded settings below. All baselines record `anatomy=0` and have no live duplication-loss series; all treatments have a live series. Raw recorded settings are separated below from values resolved during execution. Under the author-confirmed execution interpretation, soft duplication was the only additional research loss in the treatment arm; the dataset-specific standard task losses remained active in both arms, and other research-loss parameters stored in archived configurations were inert.

The checkpoint git fields are null. The February source revisions used for the scheduler and optimizer audit are therefore fingerprint-inferred from checkpoint dates and merged arguments rather than confirmed by an embedded SHA. The `repro_v4` package provides the isolated executed equation, closed-form gradient, dataset-specific weight function, five-seed manifests, and a four-check equivalence test.

**Table 7.** Dataset-level audit of the two evaluated label schemas.

| Dataset | Images | Schema classes | Images $>$ 32 boxes | Duplicate-number images |
|---|---|---|---|---|
| HITL | 595 | 32 | 0 | 0 |
| DENTEX | 634 | 32 | 2 | 16 |

**Table 8.** Execution-backed training configuration. Batch entries follow seed order 42/123/456/789/2024 and retain baseline/treatment differences.

| Setting | HITL | DENTEX |
|---|---|---|
| Task / model | segment / YOLOv11n-seg | detect / YOLOv11n |
| Standard task losses | box/segmentation/classification/DFL | box/classification/DFL |
| Configured duration | up to 100 epochs; early stopping, patience 50 | up to 100 epochs; early stopping, patience 30 |
| Completed rows, baseline | 68/86/73/83/84 | 73/75/85/77/85 |
| Completed rows, treatment | 87/92/77/92/95 | 93/68/81/100/89 |
| Training image size / rect | 1024 / true | 1024 / false |
| Seeds | 42, 123, 456, 789, 2024 | |
| Requested batch | `-1` (AutoBatch) | `-1` (AutoBatch) |
| Resolved baseline batches | 24/23/23/23/23 | 56/44/42/44/42 |
| Resolved treatment batches | 17/23/23/23/24 | 136/54/51/51/51 |
| Recorded optimizer / nominal lr0 | `auto` / 0.01 | `auto` / 0.01 |
| Resolved optimizer / learning rate | AdamW / $2.78 \times 10^{-4}$ | AdamW / $2.78 \times 10^{-4}$ |
| Resolved first-moment coefficient | 0.9 | 0.9 |
| Recorded gains (box/cls/dfl) | 7.5/0.5/1.5 | 7.5/0.5/1.5 |
| Additional treatment research loss | $\mathcal{L}_{\text{dup}}$ only | $\mathcal{L}_{\text{dup}}$ only |
| Recorded `anatomy` weight (base/treatment) | 0 / 0.5 | 0 / 0.5 |
| Effective treatment weight | 0.165 $\rightarrow$ 0.5 by ep. 20; terminal 0.525–0.975 | constant 0.165 |
| Configured ep. 99 endpoint | 1.1, not reached | not applicable |
| Fixed evaluation | 1024 px, rect=false, confidence 0.25 | |